Robust Pitch Extraction in Pathological Voice Based on Wavelet and Cepstrum

Hu Weiping[1], Wang Xiuxin[1] and Pedro Gómez[2]

[1]Guangxi Normal University, Guilin,Guangxi, P.R.China,541004
[2]Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
Email: huwp@mailbox.gxnu.edu.cn

ABSTRACT

This paper proposes a new method for pitch extraction, especially useful for pathological voice, by using the wavelet transform in the frequency domain, and disregarding the upper half signal. In this way the benefits of discrete dyadic wavelet transform are combined and the performance of the cepstral method is dramatically improved. This method can also be used for robust pitch extraction in noisy environments.

1. INTRODUCTION

The accurate determination of the pitch period or fundamental frequency of speech has been as an important research topic in speech signal processing. Classically, there are three main methods to extract the pitch which have been proposed [2-4]:

- Frequency methods such as FFT, or Cepstrum.
- Temporal methods based on the autocorrelation function such as AMDF, LPC, PPA,etc.
- Time-frequency methods such as MPDA,wavelet,ect.

Although lots of improved methods have been proposed in this direction and some of them work quite well even in noisy environments [4], for pathological voice, there is not any error free method to accurately extract the pitch so far due to the special character of degraded voice found.

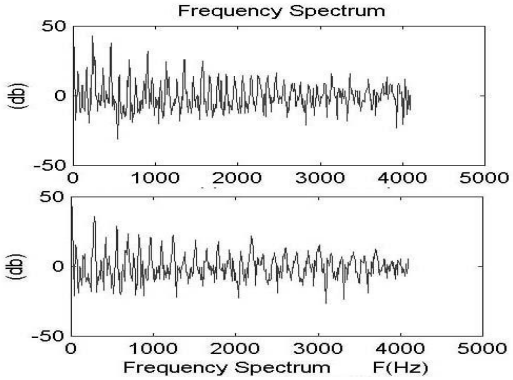

Figure1: Typical Spectrum of pathological voice.

The typical and apparent character of pathological voice is rather rough. Two main effects result from this "roughness". In the frequency domain, the periodicity is corrupted specially in high frequency bands, producing spectra similar to the those related to normal noisy speech, the other effect is that there are one or several superimposed oscillation peaks between two main harmonic peaks, as shown in Figure 1. This superimposed oscillation causes the problems found in conventional pitch extraction. In some cases the superimposed oscillation peak is so strong that it renders the traditional method unable to estimate the pitch correctly! The cepstrum (CEP) method [5] is one of the traditional methods to extract the pitch, which makes use of the spectral characteristics of speech signals. This method is able to accurately extract the pitch with little influence of the vocal tract due to the liftering process carried out. Nevertheless it can do a good job as far as clean speech signals are concerned, but it is not so efficient in noisy environments. In recent years several methods based on CEP have been proposed, some improved methods have also been designed to deal with speech signals obtained from noisy environments [2,4], but none of them can deal correctly with pathological voice, because of the second characteristic found in pathological voice. Therefore, when a log spectrum (frequency domain) of pathological voice is observed, the following two facts may be noticed: on one side, unexpected peaks appear regularly on the spectral valleys specially in the low frequency domain due to the superimposed oscillation of pathological voice. On the other side the periodicity of log spectrum is corrupted in high frequency bands due to the roughness of voice. Aiming at the two special characteristics of morbid speech, it is still possible to carry out some processing to get the periodical frequency of pathological voice accurately and correctly. In this paper, combining the benefits of the discrete dyadic wavelet transform and an improved CEP technique, a new version of CEP is proposed including three operations to remove the influence of the morbid factor. By comparing the proposed method with the conventional methods (including MCEP) [2], it may be shown that the method presented is effective not only to extract the pitch in pathological voice but also to deal with noisy speech as well.

2. PITCH DETECTION ALGORITHM

An improved CEP method for pitch extraction in noisy environments has been proposed (which will be referred to as MCEP). It will be shown that the MCEP does work quite well with noisy voice according to experimental results [2]. The flow graph of this modified CEP (MCEP) is shown in Figure 2.

The specificity of MCEP is the "clipping" performed, discarding the frequency components above 1600Hz which are corrupted by noise. This step is very useful and important to improve the MCEP performance in normal noisy voice. According to our experience working with Chinese patients, abnormal voice would be corrupted by noise above 1700 Hz [1] under a statistical point of view, and the "rough" the voice, the smaller the portion in the frequency domain that will be corrupted by noise. This was proved also by the experiment of Hajime which showed that

the best results would be obtained by removing high frequency components above 1600Hz [2]. Although the MCEP can solve the noise problem, it is useless regarding another important problem related with pathological voice, which is the treatment of superimposed oscillation peaks, which mainly take place in the lower range of frequency components.

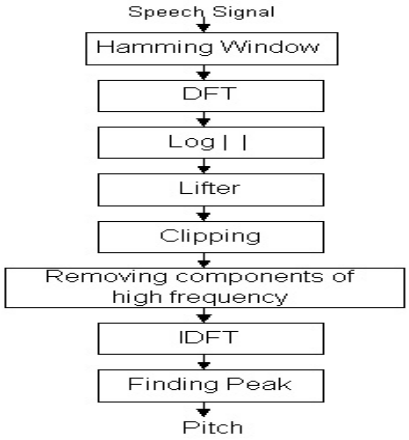

Figure 2: Flow chart of MCEP method

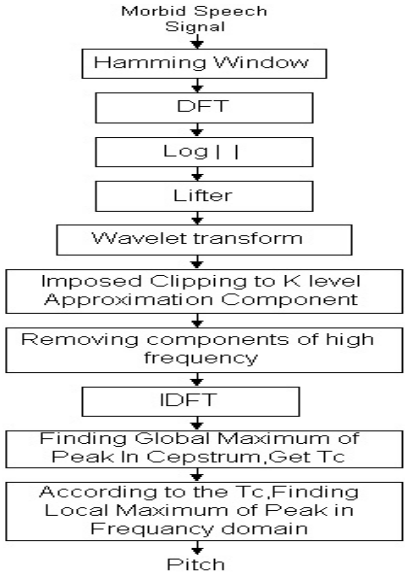

Figure 3: Flow chart of ACEP Method

By analysing the superimposed oscillation peak problem, we find that it often happens at the double, triple or quadruple harmonic of the fundamental frequency (pitch), therefore it is possible to use a low-pass filter in the frequency domain to suppress these unnecessary peaks so that the fundamental frequency component can be efficiently projected. The discrete dyadic wavelet, in essence, is a pair of low-pass and high-pass filters, and the different level of approximation components are those corresponding to low-pass filtering. Therefore when the dyadic wavelet transform is applied to the log spectrum, all what has to be done is just scrutinising every approximation component of different level and ensure that the *k-th* level (or the scale 2^k) approximation component can meet the requirement for suppressing the unnecessary peaks and projecting the fundamental component. Figure 3 shows the block diagram of the proposed method: advanced CEP (ACEP). Liftering is carried out to remove the influence of vocal tract on the spectrum. The wavelet transform is carried out to level k (or at scale 2^k) without reducing the samples of the spectra, and the value of k will depend on the sample rate of the speech signal. Regarding clipping, the *k-th* level-approximation component portion under 1600Hz is used to perform the IDFT (Inverse Discrete Fourier Transform), and instead of assigning the global maximum T_c in Cepstrum to the pitch, the local maximum according to the T_c in the frequency domain is searched and used, because the accuracy of pitch got from the frequency domain would be much more accurate than the pitch directly estimated from Cepstrum [2].

3. IMPLEMENTATION AND RESULTS

Pathological and normal voice segments of 1024 16-bit samples taken at 8192Hz are used in our experiments. All the data are obtained from a sustained phonation of vowel /a:/ during three seconds. The total number of pathological voice samples involved in the experiments is from 250 patients. Experiments include normal CEP method, MCEP method and ACEP method. Because the maximum frequency component in the samples is no more than 4096Hz and the data frame length is 1024, the $k=3$ level wavelet approximation component was used in the ACEP method. All cepstrum curves are ploted starting from point=4. In all the figures below, A_0 represents the original signal and A_2, A_3 represent level 2 and level 3 wavelet approximation components, respectively. The experiments were divided into two different sets. Experiment *A* was designed to deal with abnormal voice and experiment *B* is designed to deal with normal noisy voice under different SNR conditions.

Experiment *A*: 256 pathological voice samples were used. Results reveal that the normal CEP method failed to treat 92 of the samples, and that the MCEP method failed to treat 76 of them, but using the ACEP methods, only 12 fails in estimating the correct pitch were recorded, as seen in Table 1.

Total # of samples: *256*	Number of Fails in estimating correct Pitch
CEP	*92*
MCEP	*76*
ACEP	*12*

Table 1: Estimating the pitch in pathological voice.

Figures 4.a, 5.a, and 6.a are three typical power spectrum signals from different levels of the wavelet transform, and Figures 4.b, 5.b and 6.b are the corresponding cepstrum curves estimated using the three mentioned methods. Figure 4.a shows a typical sample X1 that the periodicity of log spectrum is heavily corrupted above 2500Hz, and the superimposed oscillation peaks can be seen in almost all the frequency domain, therefore its corresponding cepstrum curve (Fig. 4.b) suggests that normal CEP and MCEP fail to estimate the correct pitch, because it is rather difficult to find the real pitch peak, but the ACEP can work correctly just by looking for the global maximum! Similar cases are shown in Figs. 5 and 6. In Fig. 5.a, the spectrum (of typical sample X2) above 1500Hz is very noisy and the

superimposed oscillation peaks also appear in low frequency components.

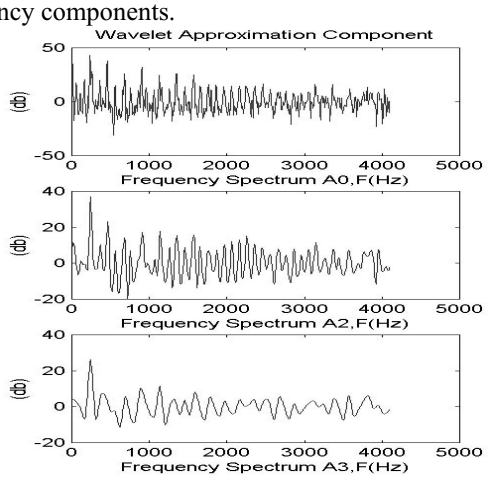

Figure 4.a: Power spectrum of typical sample X1.

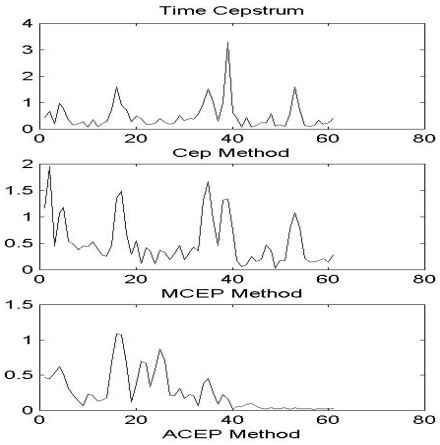

Figure 4.b: Cepstrum of typical sample X1.

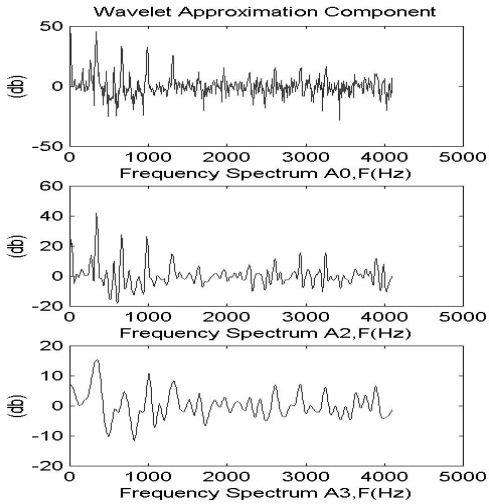

Figure 5.a: Power spectrum of typical sample X2..

The case shown in Fig. 6a typical sample X3, corresponds to a rather rough voice, because the harmonic peak can be seen below 1000 Hz in the frequency domain. The results shown in Figs. 5.b, 6.b also reveal that the ACEP can still work correctly in a case of extraordinary rough voice which the CEP and MCEP can't handle at all! By analysing the 12 cases which ACEP fail to deal with, we found that 5 of them correspond to extremely rough voice, such that the minimum periodicity limit for the ACEP to operate properly is almost reached.

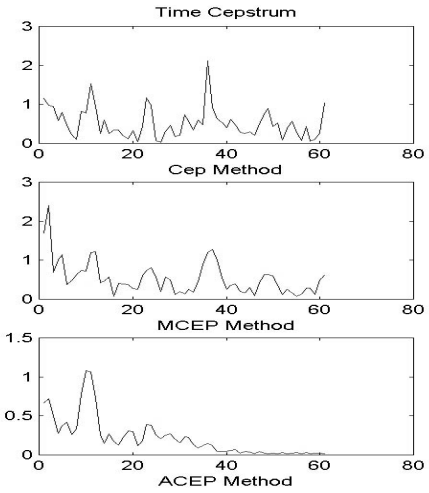

Figure 5.b: Cepstrum of typical sample X2.

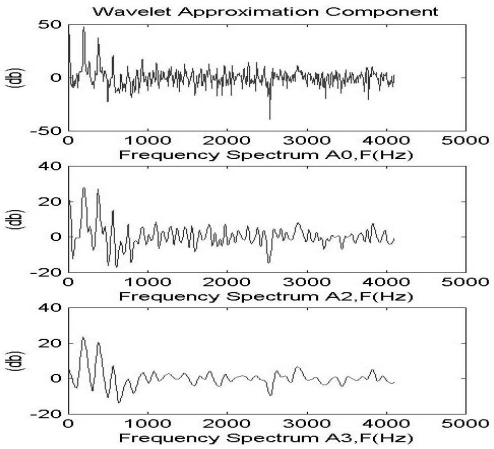

Figure 6.a: Power of typical sample X3.

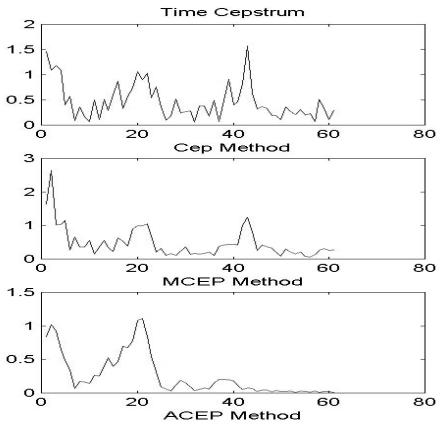

Figure 6.b: Cepstrum of typical sample X3.

The other 7 cases correspond to voice samples in which pitch is too low, less than 110 Hz. All come from male patients, therefore their corresponding frequencies are such that have been suppressed by level-3 low-pass filters. The direct result is the real pitch peak shifted to the low frequency range, thus causing the problem.

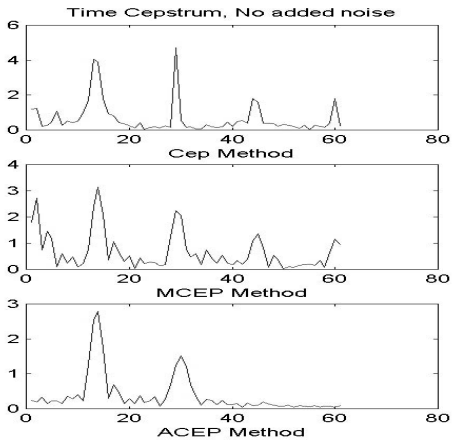

Figure 7: Cepstrum of normal sample X4

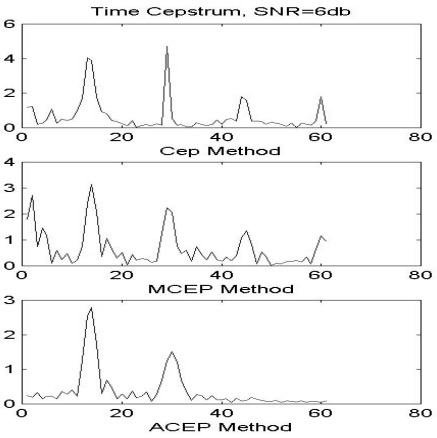

Figure 8: Cepstrum of normal sample X4

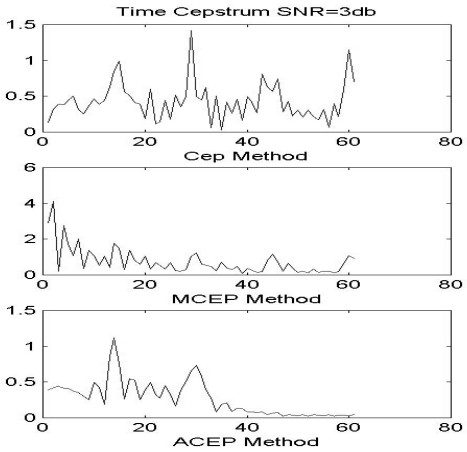

Figure 9: Cepstrum of normal sample X4

Experiment *B*: One normal sample X4 from female voice corrupted with different SNR levels of 0dB, 3dB, 6dB, and no added noise, were used in this experiment. According with what would be expected, as the levels of added noise increased, the pitch extraction performance of both CEP and MCEP decreased, and when the SNR was less than 3db, both CEP and MCEP won't work correctly while the ACEP can work correctly all the time! These results are shown in Figs. 7, 8, 9 and 10.

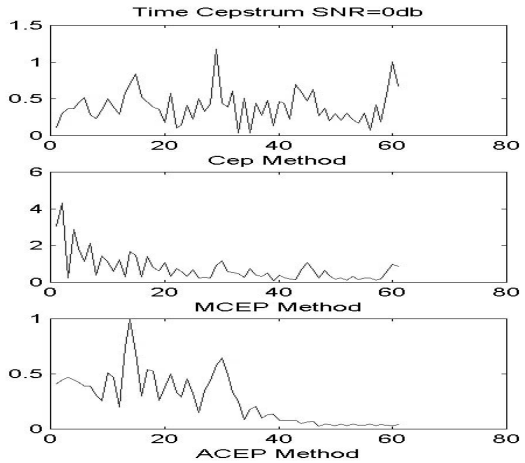

Figure 10: Cepstrum of normal sample X4

4. CONCLUSIONS

The ACEP method does exhibit a tremendous improvement in pitch extraction ability of pathological voice performing much better than CEP and MCEP in normal noisy voice, due to the benefits of wavelet low-pass filtering and clipping. But it should be noted that the ACEP presents a drawback, as it fails to deal with to very low pitch samples, regularly coming from male patients. The only approach to overcome this inconvenience is to adjust the wavelet transform levels automatically according to male/female condition, this being left as future work. Concluding, cepstral analysis based on clipping and wavelet-based low-pass filtering offers an effective and robust method not only to estimate pitch in pathological voice but also to deal with noisy speech as well.

ACKNOWLEDGMENTS

This research is being funded by the Natural Science Foundation of Guangxi Province, P. R. of China and in part with funds from the project of Programa Nacional de las Tecnologías de la Información y las Comunicaciones from the Ministry of Science and Technology of Spain No. TIC-2002-0273

REFERENCES

[1] Hu Weiping et al. "Objective evaluation of basic quality of the voice", *Journal of Audiology And Speech Pathology (China)*, Vol.6, No.4, 1998, pp. 193-195.

[2] Kobayashi, H., Shimamura, T. "A modified cepstrum method for pitch extraction"*, Proc. of 1998 IEEE Asia-Pacific Conf. on Circuits and Systems*, 24-27 Nov, 1998, pp. 299 – 302.

[3] Jianling Hu, Sheng Xu, Jian Chen, **"**A modified pitch detection algorithm**"** *IEEE Comm. Let. ,* Vol. 5, No. 2 , February 2001, pp 64 –66.

[4] Chen, S.-H.; Wang, J.-F. "Noise-robust pitch detection method using wavelet transform with aliasing compensation" ; *Proc. of the IEE Vision, Image and Signal Processing*, Vol. 149, No. 6, December 2002, pp. 327 – 334.

[5] Noll, A. M., "Cepstrum pitch determination" *J. Acoust. Soc. Amer.*, Vol. 41, No. 2, 1967, pp. 293-309.